



Otto-von-Guericke-Universität Magdeburg

Fakultät für Informatik  
Institut für Technische und Betriebliche Informationssysteme  
Prof. Dr. A. Nürnberger

Magdeburg, 16.07.10

## Prüfung Bioinformatik

Name:	Matrikelnummer:
Vorname:	Studiengang:
Blattanzahl:	
Unterschrift Student/in:	Unterschrift Aufsicht:

Aufgabe 1	Aufgabe 2	Aufgabe 3	Aufgabe 4	Summe
(von 9)	(von 13)	(von 11)	(von 10)	(von 43)

### Allgemeine Hinweise:

- Schreiben Sie auf jedes Blatt Ihren Namen, Ihre Matrikelnummer und die Seitennummer!
- Beginn und Ende einer Aufgabenlösung sind durch einen waagerechten Strich deutlich zu kennzeichnen! Ungültige Lösungen sind durchzustreichen!
- Zugelassene Hilfsmittel: ausschließlich Schreibmaterialien
- Die Klausur besteht aus 4 Aufgaben, die Bearbeitungszeit beträgt 120 Minuten.
- Bitte nicht mit grünem oder rotem Stift schreiben!

## Aufgabe 1 - Datenbanken (9 Punkte)

a) Gegeben sei eine SQL-Anfrage:

```
SELECT
  Species.Name, Est.Sequence, Est.Date,
  Laboratory.Name, Laboratory.Address
FROM
  Species, Est, Laboratory
WHERE
  Est.Species = Species.ID
  AND Est.Laboratory = Laboratory.ID
  AND Laboratory.Country = "Germany"
```

Nennen Sie für jede Relation der zugrundeliegenden relationalen Datenbank deren Namen und die für die Durchführbarkeit der SQL-Anfrage notwendigen Attribute! (2 Punkte)

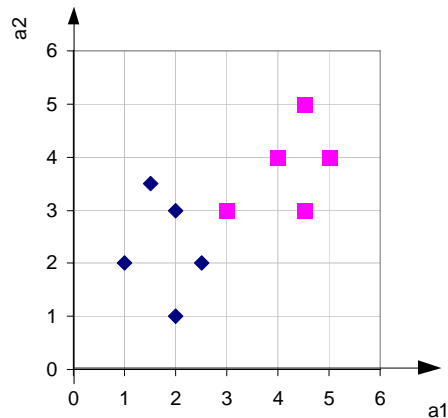
- b) Es soll eine Datenbank für die cDNA von biologischen Spezies aufgebaut werden. In dieser sollen dabei sowohl die cDNA der Spezies, als auch der bearbeitende Wissenschaftler und das Datum des Eintrages gespeichert werden. Für die Wissenschaftler soll der Name und die Kontaktadresse gespeichert werden. Für die Spezies soll ihr Name und eine Beschreibung gespeichert werden. Beachten Sie, dass für jede Spezies mehrere Einträge gemacht werden können. Jeder Wissenschaftler kann Einträge für verschiedene Spezies anlegen. Modellieren Sie diesen Sachverhalt in einem ER-Schema der 3. Normalform! (3 Punkte)
- c) Erstellen Sie für das Schema aus b) die Tabellen einer relationalen Datenbank. Kennzeichnen Sie Primär- und Fremdschlüssel! (2 Punkte)
- d) Gegeben sei eine Beschreibung der molekular-biologischer Datenbank:

```
"The database currently holds more than 200 000 sequences automatically derived from
41 cDNA libraries of four species: barley, wheat, pea and potato."
```

Welche Eigenschaft(en) nach Leser ist (sind) in dem Text beschrieben? Nennen Sie diese und beschreiben sie kurz. Beschreiben Sie außerdem weitere mögliche Ausprägungen dieser Eigenschaft(en)! (2 Punkte)

### Aufgabe 2 - Sequenzvergleiche (13 Punkte)

- Suchen Sie in der Sequenz TGDTCGTACAGTCGTAAC nach dem ersten exakten Vorkommen der Sequenz GTA ! Verwenden Sie dazu einen der besprochenen Suchalgorithmen für Zeichenketten! Wie viele Vergleiche sind notwendig? Nennen Sie alternative Suchverfahren! (3 Punkte)
- Erstellen Sie für die Sequenzen ATGTACGGTGCAT und ATGGTGCAT einen Dotplot! Welches ist das optimale Alignment? (2 Punkte)
- Berechnen Sie den Hamming- und Jaccard-Abstand (auf 4-Grammen) zwischen der Sequenz ACCTAG und CCTAGA! (2 Punkte)
- Wie groß ist der Editabstand zwischen den Sequenzen TGAACT und TTCC? Beachten Sie, dass Sie (z.B. durch Systematik oder mit einem bekannten Algorithmus) zeigen müssen, dass der von Ihnen bestimmte Abstand tatsächlich minimal ist! (2 Punkte)
- Erläutern Sie die Arbeitsweise des BLAST-Algorithmus für die Suche in einer DNA-Datenbank. (2 Punkte)
- Betrachten Sie die klassifizierten Punkte im Diagramm als Fallbasis. Klassifizieren Sie mittels des k-Nächste-Nachbarn Algorithmus (mit k=3 und dem maximalen Abstandsmaß) den Punkt (2.5; 4). (2 Punkte)



$$d_{\infty}(x, y) = \max_{i=1}^n |a_i(x) - a_i(y)|$$

### Aufgabe 3 – Expressionsdatenanalyse (11 Punkte)

- Beschreiben Sie den grundlegenden Ablauf eines Microarray-Experiments! (2 Punkte)
- Was versteht man unter MAGE-ML? Erläutern Sie die Unterschiede zwischen MIAME und MAGE-ML! (2 Punkte)
- Wieso wird eine Normalisierung der Ergebnisse eines Microarray-Experiments gebraucht? (1 Punkt)
- Führen Sie für die folgenden Messwerte eines Experiments eine Durchschnittsnormalisierung durch:

	Hyb1	Hyb2	Hyb3
Gen1	10	400	4000
Gen2	20	250	1000
Gen3	5	250	3000
Gen4	5	100	2000

(2 Punkte)

- Gegeben sind Ergebnisse einer Expressionsdatenanalyse von Gen X auf Ausprägung von gesunden und kranken Zellen, wobei bei den kranken Zellen Gen X stärker exprimiert zu sein scheint. Welche Verfahren gibt es um zu prüfen, ob die stärkere Ausprägung signifikant ist. Beschreiben Sie einen der Ansätze kurz! (2 Punkte)
- Was versteht man unter Ko-Regulation von Genen? Nennen Sie zwei Verfahren die dazu dienen Ko-Regulation von Genen festzustellen! (2 Punkte)

#### **Aufgabe 4 – Phylogenie (10 Punkte)**

- a) Welche Ansätze zur Ableitung verwandtschaftlicher Beziehungen zwischen Spezies sind Ihnen bekannt? Nennen Sie jeweils einen Beispielalgorithmus! (2 Punkte)
- b) Wann ist eine Matrix additiv? (1 Punkt)
- c) Welche der folgenden Matrizen ist eine ultrametrische Abstandsmatrix? Begründen Sie Ihre Entscheidung! (3 Punkte)

<b>M<sub>1</sub>:</b>	A	B	C	D	<b>M<sub>2</sub>:</b>	A	B	C	D	<b>M<sub>3</sub>:</b>	A	B	C	D
A	0	9	9	5	A	0	1	6	1	A	0	2	9	5
B	9	0	7	9	B	1	0	1	7	B	2	0	7	3
C	9	7	0	9	C	6	1	0	1	C	9	7	0	8
D	5	9	9	0	D	5	7	1	0	D	5	3	8	0

- d) Welche Arten von Bäumen werden von UPGMA konstruiert? Welche Anforderungen werden an den biologischen Entwicklungsprozess gestellt damit UPGMA korrekt angewendet werden kann? Beschreiben Sie diese! (2 Punkte)
- e) Wenden Sie auf die folgende Abstandsmatrix UPGMA-Verfahren an und stellen Sie den resultierenden phylogenetischen Baum grafisch dar! (2 Punkte)

	A	B	C	D	E
A	0	5	9	5	4
B	5	0	1	10	6
C	9	1	0	8	2
D	5	10	8	0	3
E	4	6	2	3	0